



(19) Europäisches Patentamt  
European Patent Office  
Office européen des brevets



(11) Publication number: 0 622 742 A1

(12)

## EUROPEAN PATENT APPLICATION

(21) Application number: 94302567.6

(51) Int. Cl.<sup>5</sup>: G06F 15/38

(22) Date of filing: 12.04.94

(30) Priority: 28.04.93 JP 102973/93

(43) Date of publication of application:  
02.11.94 Bulletin 94/44

(84) Designated Contracting States:  
DE FR GB

(71) Applicant: International Business Machines Corporation  
Old Orchard Road  
Armonk, N.Y. 10504 (US)

(72) Inventor: Uramoto,Naohiko  
1377-1-103,Kogasaka  
Machida-shi,Tokyo-to 194 (JP)

(74) Representative: Davies, Simon Robert  
I B M  
UK Intellectual Property Department  
Hursley Park  
Winchester, Hampshire SO21 2JN (GB)

### (54) Language processing system.

(57) Source sentences in the first language to be translated are entered to an input unit 12. Through the input unit 12 source sentences are entered to a polysemous word settlement unit 20 via an entered sentences analyzing unit 14. A characteristics table generation part 30 stores a characteristics table in which each meaning is preliminarily associated with each characteristic with respect to polysemous words. A characteristics record generation part 22 generates a characteristics record to compare the characteristics table and the characteristics record, thereby selecting and outputting the meaning of the polysemous word based on the characteristics in the source sentences. Each word whose ambiguity has been settled at the polysemous word settlement unit 20 is translated at a correlative translation processing unit 40 to be output to an output unit 46. Thus ambiguity of the meaning of a polysemous word in a first language having a plurality of meanings is settled so as to make source sentences in the first language correspond in the most appropriate way to the translated sentences in the second language.

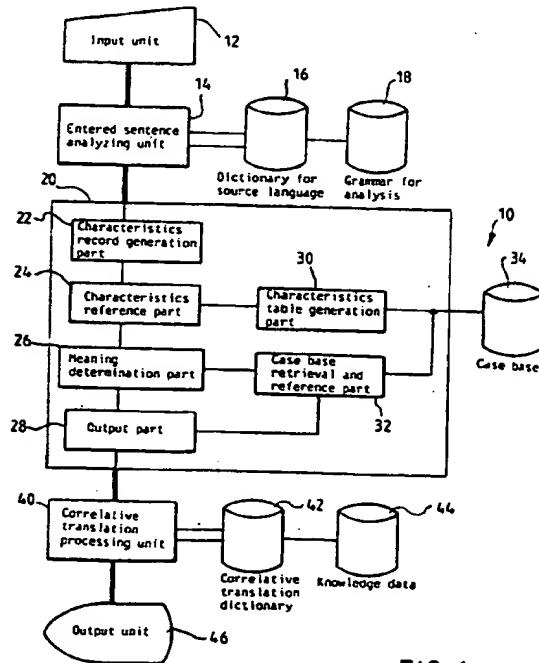


FIG. 1

The present invention relates to a language processing system, and in particular to machines such as computers used to translate source sentences in a first language into translated sentences in a second language.

A conventional machine translation apparatus is known in which a computer carries out analysis, such as morphological analysis, of the structure of source sentences expressed in a first language so as to divide the source sentences into words so that each sentence in the first language is associated with a counterpart in the second language (see JP-A 2-308370 and JP-A 3-8082 and the like).

Such machine translation apparatus performs structural analysis, such as correlative analysis, with respect to source sentences in a first language to be translated. Temporary correlation is attached to a portion of the source sentence whose meaning is ambiguous, so that the portion is translated into the second language by making use of the word order in the source sentence. With respect to ambiguous portions, information in the other portions is utilized so as to settle the ambiguity thereof to permit translation into the target language.

However it is very difficult to associate source sentences with the translation sentences on a word level when the source sentences contain words having a plurality of meanings (hereinafter referred to as polysemous words) that make the whole sentence ambiguous. Since it is necessary to perform case-by-case analysis of the usage of each word through detailed rules determined by sufficient analysis, such as syntactic analysis, it so happens that an indefinite number of translations can be associated with the polysemous word in source sentences and one meaning cannot be selected.

Accordingly there is known a method for machine translation wherein a verb conversion table is formed as a means for determining the meaning of polysemous words for the process of assigning a translation word in a second language to a polysemous verb in a first language, the verb conversion table describing translation and usage for each polysemous verb so that the table can be referred to in order to translate such a verb into a suitable meaning (JPA 62-219165).

However, in the method for machine translation using a verb conversion table which describes only verbs, information is applied to source sentences uniformly only with respect to verbs. Thus it is impossible to select the most suitable meaning for all the polysemous words contained in the analyzed sentences except for verbs. In addition, it is impossible to handle various semantic information and grammatical information dynamically in determining the meaning of each word.

In addition, a method is also known for machine translation wherein polysemous words are ordered into a database comprising sentence cases containing situations and attributes, said database being utilized to set a suitable relation between words and the meanings thereof (JP-A 3-282677). However, in such a method for specifying the meaning of words from case bases in which a plurality of meanings are generalized, the quantity of calculation required for determining the meaning of words is greatly increased.

Thus the prior art does not provide a satisfactory solution to the problem of the machine translation of source sentences into translation sentences with high accuracy, including settling ambiguity in words having plural meanings in a first language so that source sentences in a first language can be most suitably associated with the translation in a second language.

Accordingly the invention provides a language processing system for machine translation wherein source sentences in a first language which contain polysemous words having a plurality of meanings are analyzed to obtain semantic information and grammatical information, said semantic information and grammatical information being utilized to translate the source sentences into target sentences in a second language; the system being characterised by:

means for storing a characteristics table describing the relation between sufficient characteristics and meanings corresponding to said sufficient characteristics, said sufficient characteristics being at least one of either semantic information or grammatical information corresponding to a single meaning of the plural meanings of a polysemous word, the characteristics table being derived from a plurality of sentences each of which contains a polysemous word expressed in an unambiguous manner;

means for determining whether at least one of either the semantic information or the grammatical information of an analyzed sentence containing a polysemous word matches one of said sufficient characteristics; and

means for selecting the meaning of the polysemous word in accordance with the matched sufficient characteristics.

In such machine translation structures of sentences in a first language to be translated containing a polysemous word having a plurality of meanings are analyzed to determine semantic information and grammatical information, the semantic information and grammatical information being utilized to translate source sentences into sentences in a second language. Polysemous words cannot be directly translated into the second language because of the uncertainty or ambiguity in their meaning. Generally polysemous words have distinctive characteristics for each meaning that serve to distinguish their meanings from each other. In other words semantic information may indicate that the object is a material or the subject is an opponent, whilst grammatical

information may indicate that the verb is used as a transitive verb or the verb has two objects. These characteristics can be grammatical information or semantic information (or both).

The machine translation involves preliminarily obtaining the relation of sufficient characteristics to the meanings corresponding to said sufficient characteristics, said sufficient characteristics being information which is at least one of either the semantic information or grammatical information corresponding to a single meaning of the plurality of meanings of a polysemous word from a plurality of sentences each of which contains polysemous words and is construed so that one of said polysemous words is expressed in an unambiguous manner. It is determined from the obtained information whether one of either the semantic or grammatical information in a first language sentences whose structure has been analyzed matches the sufficient characteristics. When the meaning of a polysemous word corresponding to matched sufficient characteristics is selected, the ambiguity of meaning in the sentence in the first language is settled and the sentence can be easily translated into the second language. Thus the sufficient characteristics distinguishing the meaning from other meanings is determined in order to select the meaning. Consequently no high load similarity calculation needs to be performed.

It is possible to further enhance the above-described translation process by noting that when one meaning in a polysemous word has characteristics common to all the sentences corresponding to the meaning, the sentences to be entered must contain this information in common. The machine translation can therefore involve preliminarily determining the relation of the sufficient characteristics to the meaning corresponding to said sufficient characteristics and the relation of required characteristics to the meaning corresponding to said required characteristics, the required characteristics being information which is at least one of either the semantic information or the grammatical information corresponding to all the sentences with respect to a single meaning of a polysemous word. It is determined from the obtained relation whether at least one of either the semantic information or grammatical information in a first language sentences whose structure has been analyzed matches the sufficient characteristics and the required characteristics to select the meaning of a polysemous word corresponding to the matched sufficient characteristics and the matched required characteristics. Thus the meaning is selected by determining the required characteristics that exist without fail with respect to the meaning determined by the sufficient characteristics which the sentences to be translated have. Consequently, no high load similarity calculation needs to be performed and the meaning of a polysemous word can be selected with high accuracy.

The characteristics table can be used to express the relation of the sufficient characteristics to the meaning corresponding to the sufficient characteristics and the relation of the required characteristics to the meaning corresponding to the required characteristics; determination means determine whether at least one of either the semantic information or the grammatical information of the analyzed sentences matches the sufficient characteristics and the required characteristics; and selection means select the meaning of a polysemous word corresponding to the matched sufficient characteristics.

As a still further enhancement, the machine translation may involve preliminarily determining the relation of the sufficient characteristics to the meaning of the sufficient characteristics, the relation of the required characteristics to the meaning of the required characteristics and the relation of the dependent characteristics to the meaning of the dependent characteristics, the dependent characteristics being information which is at least one of either the semantic information or grammatical information corresponding to two or more meanings of the polysemous words. From the obtained relation it may be determined whether at least one of either the semantic information or grammatical information of first language sentences whose syntactic structures have been analyzed matches the sufficient characteristics and required characteristics, and similarity calculated based on sentences included in the scope of the meaning corresponding to the dependent characteristics when at least one of either the semantic information or grammatical information does not match the sufficient characteristics and the required characteristics, and selecting the meaning of the polysemous word from the result of the calculation. Typically, a meaning is selected which has a high similarity. Consequently, the meaning of the polysemous word can be determined when the first language to be translated does not have distinctive features such as sufficient characteristics or required characteristics.

Thus preferably the system includes means for determining whether at least one of either the semantic information or grammatical information of the analyzed sentences matches the sufficient characteristics and required characteristics by referring to the characteristics table, and calculation means for calculating a similarity based on sentences included in the scope of the meaning corresponding to the dependent characteristics when the determination means determines at least one with sufficient characteristics and required characteristics, and a selection means for selecting the meaning of the polysemous word based on the result of calculation.

Various embodiments of the invention will now be described in detail by way of example, with reference to the following drawings:

Figure 1 is a block view showing the general construction of a machine translation system:

Figure 2 is a flowchart showing the main routine of the machine translation apparatus;

Figure 3 is a flowchart showing the flow of the routine for generating the characteristics table; and

Figure 4 is a flowchart showing the flow of the routine for processing the settlement of polysemous word.

This first embodiment described utilizes English as a first language to be translated and Japanese is set as the second language into which the first language is to be translated, thereby translating source sentences in English into Japanese, although the invention is equally applicable to any languages.

As shown in Figure 1, a machine translation apparatus 10 comprises an input unit 12 such as user terminals (EG keyboards), an input sentence analyzing unit 14, polysemous word settlement unit 20, a correlative translating unit 40, and an output unit 46. To the input unit 12 is connected an input sentence analyzing apparatus 14 to which source sentences from the input unit 12 such as user terminals are input. To the input sentence analyzing apparatus 14 are connected original dictionary memory 16 storing a dictionary for an original language and a grammar memory 18 for storing grammar for analysis. Note that source sentences entered into the above input sentence analyzing apparatus 14 are not limited to verbiage, but may for example be symbols expressing language output by a sound recognition unit for recognizing spoken language.

The input sentence analyzing unit 14 is connected to a polysemous word settlement unit 20. Source sentences in the first language, after being subjected to syntactic analysis with the input sentence analyzing unit 14, are entered to the polysemous word settlement unit 20.

The polysemous word settlement unit 20 is a unit for selecting the meaning of a polysemous word which is ambiguous (i.e. has a plurality of possible meanings). The unit 20 comprises a characteristics table generating part 22, a characteristics referring part 24, a word meaning determining part 26, an output part 28, and a characteristics table generating part 30 and the case base retrieving and referring part 32. The polysemous word settlement unit 20 provides a case base 34. The case base 34 is connected to the characteristics table generating part 30 and the case base retrieving and referring part 32. The details of the construction of the unit 20 will be explained later. Each word whose meaning has been settled in the polysemous word settlement unit 20 is output to the correlative translation unit 40.

This case base 34 stores a plurality of cases 50<sub>i</sub> (i = 1, 2, ..., n, n represents the total number of cases) whose ambiguity in the meaning has been settled as shown in the following (source sentences) as a data base in which original English sentences are recorded as case 50<sub>i</sub>.

#### [Source sentences]

We took a cup of coffee. (50<sub>1</sub>)

The army took the fort. (50<sub>2</sub>)

Take her another cup of tea. (50<sub>3</sub>)

She took a gift from him. (50<sub>4</sub>)

To the correlative translation unit 40 is connected a correlative translation information memory 44 in which case data like idiomatic expressions as information and knowledge data as the meaning are stored via a correlative translation dictionary 42. The correlative translation dictionary 42 stores a translation word in the second language with respect to the first language. Furthermore, correlative information memory 44 may store generative grammar and generative rules in the second language.

As shown in Figure 2, when an operator enters a sentence to be translated with an input unit 12 (like a keyboard), the input unit 12 outputs this sentence to the input sentence analyzing unit 14. The input sentence analyzing unit 14 performs syntactic analysis of entered sentences (at step 100). Upon termination of syntactic analysis, the input sentence analyzing unit 14 outputs entered sentences and the result of syntactic analysis to the polysemous word settlement unit 20. The polysemous word settlement unit 20 selects one meaning of the polysemous word by referring to the characteristics table (detailed hereinbelow) preliminarily stored based on entered sentences and the result of analysis (at step 200). Entered sentences whose meaning is determined are output to the correlative translation processing unit 40. The correlative translation processing unit 40 generates translation sentences by selecting correlative translation of each word in the entered sentences having polysemous words whose ambiguity has been settled (at step 300).

A characteristics table 56k (k = 1, 2, ..., n, n represents the number of polysemous words), which will be used in selecting the meaning of polysemous words, will be explained. The characteristics table generating part 30 of the polysemous word settlement unit 20 effects the routine of the characteristics table generation processing shown in Figure 3 under the instruction of the operator. The routine may be effected at power-up time. The characteristics table may be stored in the memory of the characteristics table generating part 30, not shown in the drawings, until the content of the case base 34 is converted.

In the routine for generating the characteristics table shown in Figure 3, the process proceeds to step 402

to read the recorded case 50i from case base 34. Incidentally in reading case 50i, cases in which semantic information and grammatical information appeared two or more times in an overlapping manner are extracted so as to improve the certainty of the generated characteristics table (hereinafter referred to as characteristics). In addition, at step 402 with respect to cases that have been read, pronouns such as he, she, etc., in the above source sentences are simultaneously substituted by concept words such as person or the like as Conversion A shown hereinbelow.

[Conversion A]

10. person took a cup of coffee. (50<sub>1</sub>)  
 The army took the fort. (50<sub>2</sub>)  
 Take person another cup of tea. (50<sub>3</sub>)  
person took a gift from him. (50<sub>4</sub>)

15. At the subsequent step 404, characteristics record 54i ( $1 \leq i \leq n$ ) comprising a plurality of characteristics stored in a characteristics field of one case is generated and stored by matching the characteristics of read cases with the characteristics field 52j ( $j = 1, 2, \dots, m$ , m represents the number of characteristics that can be included in one case).

In this embodiment, the following five characteristics are used as characteristics of cases stored in each characteristics of characteristics record:

20. 1. words at the position of the subject (subject)  
 2. words at the position of the object (object)  
 3. prepositions (preposition)  
 4. words accompanying prepositions (noun with preposition)  
 5. sentence patterns (pattern)

25. Consequently in this embodiment, characteristics records 54i are stored in the order of characteristics fields 52<sub>1</sub>, 52<sub>2</sub>, 52<sub>3</sub>, 52<sub>4</sub>, and 52<sub>5</sub> like characteristics record format shown hereinbelow.

[Characteristics record format]

30. 54i = (subject), (object), (preposition),  
 52<sub>1</sub>            52<sub>2</sub>            52<sub>3</sub>  
 (noun with preposition), (pattern)  
 35.                52<sub>4</sub>            52<sub>5</sub>

40. Incidentally the number of the above characteristics is not restricted to 5, and it may be 4 or less or 6 or more. As an example of 6 or more characteristics, the following may also be used: any synonyms of words at the position of the subject or a group of words corresponding to the upper and lower concepts thereof, synonyms of words at the position of objects or a group of words corresponding to the upper and lower concepts thereof, words at the position of the second object, synonyms of words at the position of the second object or a group of words corresponding to the upper and lower concepts thereof, and synonyms of prepositions or a group of words corresponding to the upper and lower concept thereof and a group of co-appearance words.

45. At the subsequent step 406, a determination is made as to whether the generation of characteristics record 54i is completed with respect to all cases 50<sub>1</sub> to 50n recorded in case base 34. Upon completion of all the characteristics record 54i with respect to all the cases 50<sub>1</sub> to 50n, the process proceeds to step 408. At step 408, one word is set which becomes an object for generating the characteristics table 56k. The process proceeds to step 410. Hereinbelow will be detailed an example that can be initiated from setting the word "take."

50. At step 410 one characteristics record 54i including the set word "take" is read to proceed to step 412. At step 412, a characteristics table 56, concerning the word "take" is generated and stored by registering the characteristics of the read characteristics record 54i. In other words, since the word "take" in one case has a single meaning, each characteristics is registered in the characteristics table 56, wherein the meaning of the word take in one case is associated with the above five characteristics. Consequently when the characteristics are registered, the characteristic of the synonymous characteristics record is registered at a portion corresponding to the identical meaning. For example, the word "take" includes the following 10 meanings (sense):

[Meanings (sense)]

- 1. occupy
- 2. grasp
- 5      3. make picture
- 4. choose
- 5. carry
- 6. carry person
- 7. use
- 10     8. receive
- 9. pay
- 10. drink

At step 414 a determination is made as to whether the characteristics with respect to all the cases 50i including the word take recorded in case base 34 are registered in the characteristics table. Upon completion of the registration of the characteristics of all the cases 50<sub>i</sub> to 50n, the process proceeds to step 416.

An example of the characteristics table generated as shown above with respect to the word "take" is shown in Table 1.

20

25

30

35

40

45

50

55

[Table 1]

Characteristics Table for the English Verb Take						
	sense/feature	1. subject	2. object	3. prep.	4. noun with prep.	5. pattern
5	1. occupy	person (all) *enemy *army	*animal *castle	by (2) in	*surprise *battle	<u>SVO</u>
10	2. grasp	person (all)	*arm *hand	by (1)		<u>SVO</u>
15	3. make picture	person (all)	*photograph	at		<u>SVO</u>
20	4. choose					<u>SVO</u>
25	5. carry	person (all)	*person (6)	to (7) with	*hospital person (8)	<u>SVO</u>
30	6. carry person	person (all)	*person (5)		of	* <u>SVOO</u>
35	7. use vehicle	person (all)	*car *bus	to (5)	*home	<u>SVO</u>
40	8. receive	person (all)		*from	person (6)	<u>SVO</u>
45	9. pay	person (all)	*flat			<u>SVO</u>
50	10. drink eat	person (all)	*coffee *medecine *breath	of with		<u>SVO</u>

At step 416 each characteristic registered in characteristics table 56<sub>1</sub> is classified into sufficient characteristic 62, required characteristic 64, and dependent characteristic 66.

A sufficient characteristic 62 means those characteristics, each of which is registered in characteristics table 56<sub>1</sub>, which do not appear at all with other meanings. For example, since the word "person" appears more than one meaning as shown in Table 1, it cannot become a sufficient characteristic 62. The word "army" which does appear in only one meaning becomes a sufficient characteristic.

In addition, required characteristics 64 means those characteristics, each of which is registered in the characteristics table 56<sub>1</sub>, which are common to all the cases that can be associated with the meaning. For example, when the sentence patterns are referred to with respect to all the cases of the word "take" (Table 1), the word "take" is utilized as a transitive verb. The sentence patterns used with a transitive verb become a required characteristic 64.

Additionally, dependent characteristics 66 are those characteristics, each of which is registered in the characteristics table 56<sub>1</sub>, which appear simultaneously in other meanings. For example, the word "person" appears in other meanings of the word "take," it becomes a dependent characteristics 66.

Upon completion of the above classification of characteristics, at step 418 the dependency relation and

the frequency of appearance are added to the characteristics record classified into each group. In other words, characteristics that are neither sufficient characteristics 62 nor required characteristics 64 have a relation of dependency with other cases added on. The relation of dependency is not determined from the above characteristics table 56<sub>1</sub>, but is determined from the characteristics record of the cases. Consequently the relation of dependency can be considered in characteristics that appear only once in a case.

In the characteristics table 56<sub>1</sub>, shown in the above Table 1, symbol \* is added to words that constitute sufficient characteristics 62, and words that constitute required characteristics 64 are underlined, so that the above characteristics can be classified. In addition, no symbol is added to the dependent characteristics 66. However, the relation of dependency is shown in Table 1 by adding a subscript in a round bracket. In Table 1 numbers in the bracket correspond to the number of the associated meaning. The subscript (1-9) designates that there is the relationship of dependency between the number 1 and the number 9. The subscript (all) designates the relationship of dependency between all the meanings.

Incidentally (the description is omitted here) the degree designating the frequency is associated with every characteristic. At subsequent step 420, the presence of the polysemous words is determined and the presence of words that become an object and generate the characteristics table is determined. When a polysemous word still exists, the process returns to step 408. By repeating the above processing, a characteristics table 56<sub>1</sub> with respect to each polysemous word is generated. When generation of the characteristics table is completed with respect to all the polysemous words, the routine is completed.

The characteristics table is generated from the case base 34 storing source sentences. The table expresses the characteristics corresponding to the extracted meaning with respect to each case of one word. Thus the characteristics in the case can be extracted as the characteristics for differentiating one meaning from the other.

Furthermore, the characteristics table that has been preliminarily prepared may be stored as the above described characteristics table. In addition, the above characteristics table is not restricted to being generated for each word, but information with respect to a plurality of polysemous words may be stored in the characteristics table. One characteristics table may be divided so that different polysemous words are associated with different regions.

Subsequently step 200 will be detailed hereinbelow. As shown in Figure 4, step 200 is a subroutine in which the above generated characteristics table is used to settle polysemous words to select the meaning thereof. In other words, the characteristics are extracted from the entered sentences at step 200. When the characteristics satisfy the sufficient characteristics 62 and the required characteristics 64 of one meaning, the meaning is defined as that of a polysemous word.

To simplify the explanation, a case will be explained in which the following entered sentence 1 is used as an example of a sentence to be entered to select the meaning of the polysemous word "take" that has been used to generate the characteristics table 56<sub>1</sub>.

[Entered sentence 1]

He took the cab to the station.

When the routine is effected, the process proceeds to step 202 to generate the characteristics record 55 of the sentence entered after the completion of the analysis of sentences in the same manner as the characteristics record generation part 22 generates the above characteristics table. In this case, information that "he" is a noun and the subject is entered to the characteristics record generation part 22 when the entered sentence analyzing unit 14 analyzes the syntax.

55 = (person), (cab), (to), (station), (SVO)

At subsequent step 204, the characteristics that match the characteristics record 55 are sought by referring to the above characteristics table 56<sub>1</sub>, with respect to the characteristics record 55 generated at the characteristics referring part 24. The relation between each of the characteristics of this generated characteristics record 55 and the above characteristics table 56<sub>1</sub> are shown below in Table 2.

50

[Table 2]

Kind of characteristics	Characteristics	Classification of characteristics	Associated meaning
5 object	person	dependent characteristics	1 - 10
	cab	sufficient characteristics	7
	to	dependent characteristics	5, 7
10 noun with prep.	station	sufficient characteristics	7
	transitive verb (SVO)	required characteristics	1 - 10

A meaning differentiation part 26 selects the meaning of polysemous words based on the characteristics that match the characteristics of the characteristics table 56, determined in the above process and the characteristics of the characteristics record 55. In other words, step 206 determines whether the characteristics of the generated characteristics record 55 has the sufficient characteristics 62 of the characteristics table 56. In this case, owing to the presence of the sufficient characteristics 62 as shown in Table 2, an affirmative determination is made, and processing proceeds to step 208.

The subsequent step 208 determines whether the characteristics record 55 satisfies the required characteristics 64 of the meaning selected by the sufficient characteristics 62. In this case, since the required characteristics condition is satisfied as shown in Table 2, the process proceeds to step 210. At step 210, the presence of any residual meaning is determined. When residual meaning is absent, the meaning having the sufficient characteristics 62 and satisfying the required characteristics 64 is selected. The selected meaning is output to the correlative translation processing unit 40 at step 218, thereby terminating the routine.

Consequently the meaning determination part 26 selects the 7th meaning of the characteristics table 56, as the meaning of the word "take." The meaning is output from the output part 28 to the correlative translation processing unit 40. Thus, since the ambiguity or uncertainty of the entered sentence 1 is settled with respect to the meaning of the polysemous word, the correlative translation processing unit 40 generates and outputs the translation sentence corresponding to the source sentence as shown in the following translation sentence 1.

#### [Translation sentence 1]

35 Kare wa eki made takushii o tsukatta.

In some entered sentences the characteristics thereof may not match one of the sufficient characteristics 62 and the required characteristics 64. As an example of such a case, entered sentence 2 will be detailed hereinbelow.

#### 40 [Entered sentence 2]

He took the building.

The relation of each characteristic of the generated characteristics record 55 and the above characteristics table 56, is shown in Table 3 with respect to entered sentence 2 in the same manner as shown above.

[Table 3]

Kind of characteristics	Characteristics	Classification of characteristics	Associated meaning
50 object	person	dependent characteristics	1 - 10
	building	dependent characteristics	1, 9
	none		
55 noun with prep.	none		
	transitive verb (SVO)	required characteristics	1 - 10

As shown in Table 3, since entered sentence 2 does not contain sufficient characteristics 62, the meaning

of the word "take" cannot be selected without considering a range of meanings. When the characteristics of entered sentence does not contain sufficient characteristics 62 (a negative determination is made at step 206), or when the characteristics of the entered sentence fulfills the sufficient characteristics 62, but does not satisfy the required characteristics 64 (a negative affirmation is made at step 208), the process proceeds to step 212

- 5 to calculate the similarity between cases having the relation of dependency. In the example of this entered sentence 2, the 4th characteristic (object) having as common characteristics the relation of each characteristics of generated characteristics record 55 and the above characteristics table 56, is selected to specify the case for calculating the similarity. The similarity is calculated between the characteristics record 55 of entered sentence 2 and the characteristics record of each case of the 1st and the 9th meaning in the characteristics table 56.
- 10 The similarity calculation can be determined by calculating an agreement between characteristics records.

At subsequent step 214, the characteristic having the highest frequency is selected as the characteristic of the polysemous word in the entered sentence based on the calculated similarity. Then the process proceeds to step 218. The meaning corresponding to this characteristic is output to the correlative translation processing unit 40 and the routine is terminated.

- 15 Consequently, when the meaning of the word "take" is highly similar to the first meaning on the characteristics table 56, the first meaning (occupy) is selected. The correlative translation processing unit 40 generates the corresponding translation as shown in the following translation sentence 2.

[Translation sentence 2]

20

Kare wa tatemono o senryoshita.

- In addition, in some entered sentences, the characteristics the entered sentence contains have a plurality of meanings corresponding to the sufficient characteristics 62 and the required characteristics 64, in which case the meaning of the word "take" cannot be selected. In such a case, the characteristics of the entered sentence satisfy the sufficient characteristics and the required characteristics. When a residual meaning exists (a negative affirmative determination is made at step 210), the process proceeds to step 216. At step 216, the meaning is elected which matches the largest number of meanings in the above characteristics table. Furthermore, the frequency of characteristics is added up, and the meaning having the highest frequency may be selected. At subsequent step 218, the selected meaning is output to the correlative translation processing unit 40. This terminates the routine.

30 As described above, a characteristics table is generated which expresses the relation of each meaning of a polysemous word selected from a plurality of cases to the characteristics corresponding to the meaning. In this characteristics table, the characteristics of each meaning the polysemous word has are classified into sufficient characteristics that do not appear for other meanings, required characteristics that the entered sentence must satisfy, and dependency characteristics which relate to a plurality of meanings.

- 35 Consequently when the characteristics corresponding to the sufficient characteristics included in the characteristics table are present in the process of selecting the meaning of the polysemous word in an entered sentence, the meaning of the polysemous word can be swiftly selected without calculating similarity, which requires a large amount of calculation from the case base. By requiring that the required characteristics must be satisfied by the entered sentences in addition to the sufficient characteristics, the meaning of the polysemous word can be selected with high certainty.

40 In addition, when sufficient characteristics and the required characteristics are not available, the similarity is calculated and the meaning of the polysemous word can be selected by considering cases that are not reflected on the characteristics table. In such a case, when dependent characteristics are present, the similarity between associated meanings is calculated by referring to the dependent characteristics. Thus the amount of calculation decreases and the processing for selecting the meaning of the polysemous word can be effected in a much shorter time.

- 45 Thus, in this embodiment information such as semantic information and grammatical information is grasped as an equivalent characteristic. Consequently different kinds of information are used with respect to each meaning of a polysemous word to settle the ambiguity thereof.

50 In the above embodiment, English is used as the first language, but the present invention is not restricted to English and may be applied to any other language, using a characteristics table generated in the same manner as described above.

- 55 Next, French is used as another first language to illustrate settling an ambiguity in the French language. Since the approach is basically the same as for the English first language described above, a detailed description of the common aspects is omitted, and only differences will be described.

At the outset a characteristics table 56F<sub>L</sub> (L = 1, 2, ..., n, n represents the number of polysemous words) to utilize for settling ambiguities or uncertainty in meanings in the French language is generated by the charac-

teristics table generation part 30 in the same manner as the above. In other words, the characteristics table generation part 30 utilises the characteristics table processing routine shown in Figure 3 to read case 50i from the case base 34. The read cases are associated with the characteristics field 52j ( $1 \leq j \leq 5$ ) to generate and store a characteristics record 54i ( $1 \leq i \leq n$ ) comprising a plurality of characteristics stored in the characteristics field of one case (steps 402 and 404). After generating this characteristics record 54i with respect to all the cases 50<sub>i</sub> to 50n recorded in the case base 34, one word is identified to become the object for which the characteristics table 56F<sub>L</sub> is generated (steps 406 and 408). An example of setting the word "prendre" will be detailed hereinbelow.

A characteristics table 56F, divided according to the various meanings associated with the word "prendre" is generated and stored by reading each characteristics record 54i including the set word "prendre" and registering the characteristics of the characteristics record 54i (steps 410, 412). The word "prendre" includes the following 10 meanings.

[Meaning (sense)]

- 15      1. occuper
- 2. saiser
- 3. photographie
- 4. chosir
- 20      5. emmener
- 6. monter
- 7. recevir
- 8. payer
- 9. manger
- 25      10. succes

In the next step, after registering the characteristics of all the cases 50i including the word "prendre" recorded in the case base 34, each of the registered characteristics is classified into sufficient characteristics 62, required characteristics 64, and dependent characteristics 66 as described above, to which is added the relationship of dependency and the frequency of appearance with respect to each of the classified records (step 414 through 418).

Thus the case base 34 which stored the source sentences in French generates the characteristics table expressing the characteristics corresponding to the meaning by extracting the characteristics of each case with respect to one word. Consequently, the characteristics in the case are extracted as a characteristic for differentiating one meaning from the others.

35      An example of the characteristics table generated as shown above with respect to the word "prendre" is shown in Table 4.

40

45

50

55

[Table 4]

Characteristics Table for the French Verb "prendre"						5
	sens/trait	1. subject	2. object	3. prep.	4. nom with prep.	sens/trait
10	1. occuper	personne (1-9)	*animal *chateau	par en	*surprise *bataille	SVO
		*ennemi				
		*armee				
15	2. saisir	personne (1-9)	*arme *main	par		SVO
20	3. photographie	personne (1-9)	*photo	a		SVO
25	4. choisir					SVO
30	5. emmener	personne (1-9)	*personne	a avec	*hospital personne	SVO
35	6. monter	personne (1-9)	*autobus *voiture	to	*maison	SVO
40	7. recevir	personne (1-9)		depuis	personne	SVO
45	8. payer	personne (1-9)	*plat			SVO
50	9. manger	personne (1-9)	*cafe *medicament *air	de avec		SVO
10. succes						* SV

The characteristics table generated as shown above is used to perform the processing for settling the ambiguity of a polysemous word in French in accordance with the flowchart shown in Figure 4 in the same manner as described above for English.

Another example involving settling the ambiguity of the meaning of a Japanese word by using Japanese as the first language will now be described. Again, the approach is basically the same as described above, so that a detailed description of common parts is omitted and only differing parts will be detailed.

The characteristics table generation part 30 in the polysemous word settling unit 20 outlined above generates the characteristics table  $56J_M$  ( $M = 1, 2, \dots, n$ ,  $n$  represents the number of polysemous words) used for settling the meaning of polysemous word in Japanese. The characteristics table  $56J_M$  includes the following 10 [meanings/characteristics] in an example of the word "toru."

#### [Meaning/characteristic]

- 55 1. senryosuru
- 2. tsukamu
- 3. satsueisuru

4. erabu  
 5. saiyoosure  
 6. noru  
 7. uketoru  
 5 8. kane o harau  
 9. inshokusuru  
 10. maneku

An example of the characteristics table 56J1 thus generated with respect to the word "toru" is shown in the following Table 5.

10

[Table 5]

Characteristics Table for the Japanese Verb "Toru"					
meaning/characteristics	1. subject	2. object	3. joshi	4. noun with joshi	5. pattern
1. senryosuru toraeru	hito (1-9)  *teki  *guntai	*dobutsu  *shiro	ni  de	*totsuzen  *tatakai	SVO
2. tsukamu	hito (1-9)	*ude  *te	de		SVO
3. satsueisuru	hito (1-9)	*shashin	de		SVO
4. erabu					SVO
5. saiyoosure					SVO
6. noru	hito (1-9)	*basu  *kuruma	he	*ie	SVO
7. uketoru	hito (1-9)		kara	*hito	SVO
8. kane o harau	<u>hito (1-9)</u>	*ie			SVO
9. inshokusuru	hito (1-9)	*kohi  *kusuri	wo		SVO
40 10. maneku		*shisho	wo		SVO

The characteristics table generated as shown above is used to perform the processing for settling the ambiguity of polysemous words in Japanese in accordance with the flowchart shown in Figure 4 in the same manner as the described above.

45 Thus as described above, the machine translation uses semantic information and grammatical information from a plurality of source sentences in the first language to enable selection of the meaning of a polysemous word from the sufficient characteristics and the information of the source sentences to be translated. This approach has the advantage of swiftly selecting the meaning of a polysemous word.

50 In addition, when the analyzed sentences have information corresponding to sufficient characteristics, it is also possible to verify that the meaning has the required characteristics of information associated with all the sentences, thereby ensuring the selection of meaning of the polysemous word with high accuracy.

Even when the sufficient characteristics and the required characteristics are inadequate, the meaning of the polysemous word can be selected through calculation of the similarity by using cases in the first language. Thus advantageously the meaning of a polysemous word can be selected according to the essential meaning thereof.

55 The characteristics table is used to express the relation of information and the meaning corresponding to said information by using information which is semantic information and information on sentence patterns pos-

sessed by a plurality of sourc sentences in the first languag . This offers the advantage of s ttling the ambiguity of a polysemous word with an extremely small number of rules without calculating the similarity with respect to all the sentences and cases.

5

## Claims

1. A language processing system for machine translation wherein source sentences in a first language which contain polysemous words having a plurality of meanings are analyzed to obtain semantic information and grammatical information, said semantic information and grammatical information being utilized to translate the source sentences into target sentences in a second language, the system being characterised by:
  - means for storing a characteristics table (56) describing the relation between sufficient characteristics and meanings corresponding to said sufficient characteristics, said sufficient characteristics being at least one of either semantic information or grammatical information corresponding to a single meaning of the plural meanings of a polysemous word, the characteristics table being derived from a plurality of sentences (34) each of which contains a polysemous word expressed in an unambiguous manner;
  - means for determining whether at least one of either the semantic information or the grammatical information of an analyzed sentence containing a polysemous word matches one of said sufficient characteristics; and
  - means for selecting the meaning of the polysemous word in accordance with the matched sufficient characteristics.
2. The language processing system of claim 1, wherein the characteristics table further describes the relation between required characteristics and the meaning corresponding to said required characteristics, said required characteristics being at least one of semantic information and grammatical information corresponding to all the sentences associated with one meaning of the polysemous word, said system further comprising means for determining whether at least one of either the semantic information or the grammatical information of an analyzed sentence containing a polysemous word matches one of said required characteristics, wherein the meaning of the polysemous word is selected in accordance with the matched sufficient characteristics and matched required characteristics.
3. The language processing system of claim 2, wherein the characteristics table further describes the relation between dependent characteristics and the meaning corresponding to said dependent characteristics, said dependent characteristics being information which is at least one of either the semantic information or grammatical information corresponding to two or more meanings of said polysemous word; said system further comprising means for calculating a similarity between one of either the semantic information or the grammatical information and said sufficient characteristics and required characteristics based on said sentences included in the scope of the meaning corresponding to said dependent characteristics when at least one of either the semantic information or grammatical information does not match said sufficient and required characteristics; and means for selecting the meaning of said polysemous word based on the result of the calculation.
4. A method of machine translation wherein source sentences in a first language which contain polysemous words having a plurality of meanings are analyzed to obtain semantic information and grammatical information, said semantic information and grammatical information being utilized to translate the source sentences into target sentences in a second language, the method being characterised by the steps of:
  - storing a characteristics table (56) describing the relation between sufficient characteristics and meanings corresponding to said sufficient characteristics, said sufficient characteristics being at least one of either semantic information or grammatical information corresponding to a single meaning of the plural meanings of a polysemous word, the characteristics table being derived from a plurality of sentences (34) each of which contains a polysemous word expressed in an unambiguous manner;
  - determining whether at least one of either the semantic information or the grammatical information of an analyzed sentence containing a polysemous word matches one of said sufficient characteristics; and
  - selecting the meaning of the polysemous word in accordanc with the matched sufficient characteristics.

5. The method of claim 4, wherein the characteristics table further describes the relation between required characteristics and the meaning corresponding to said required characteristics, said required characteristics being at least one of semantic information and grammatical information corresponding to all the sentences associated with one meaning of the polysemous word, said method further comprising the steps of determining whether at least one of either the semantic information or the grammatical information of an analyzed sentence containing a polysemous word matches one of said required characteristics, and selecting the meaning of the polysemous word in accordance with the matched sufficient characteristics and matched required characteristics.
10. 6. The method of claim 5, wherein the characteristics table further describes the relation between dependent characteristics and the meaning corresponding to said dependent characteristics, said dependent characteristics being information which is at least one of either the semantic information or grammatical information corresponding to two or more meanings of said polysemous word, the method further comprising the steps of calculating a similarity between one of either the semantic information or the grammatical information and said sufficient characteristics and required characteristics based on said sentences included in the scope of the meaning corresponding to said dependent characteristics when at least one of either the semantic information or grammatical information does not match said sufficient and required characteristics; and selecting the meaning of said polysemous word based on the result of the calculation.

20

25

30

35

40

45

50

55

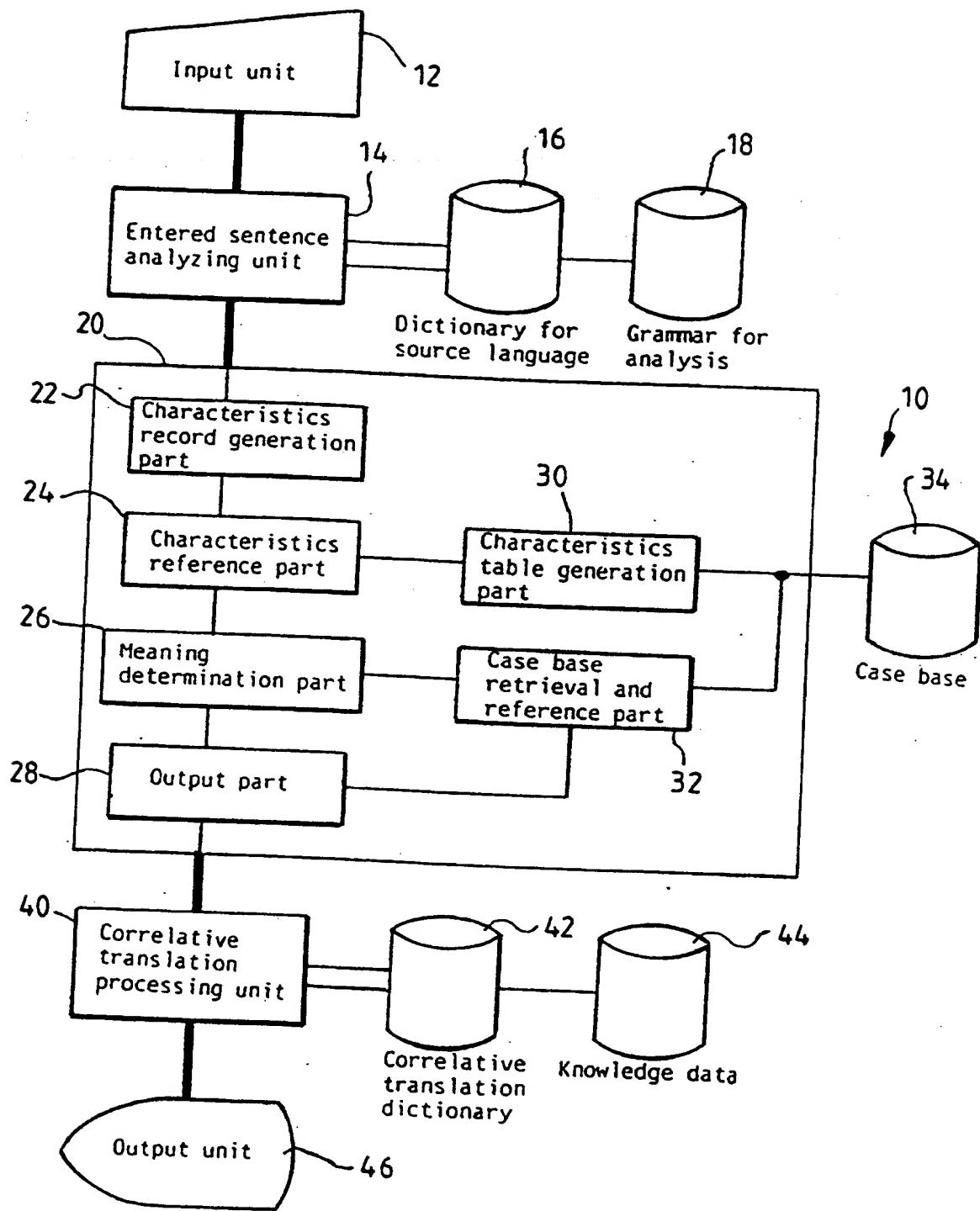


FIG. 1

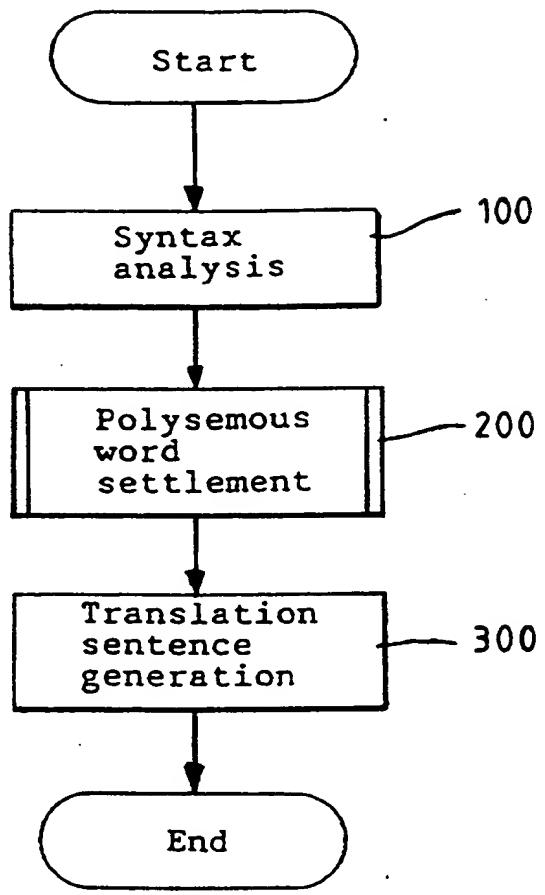
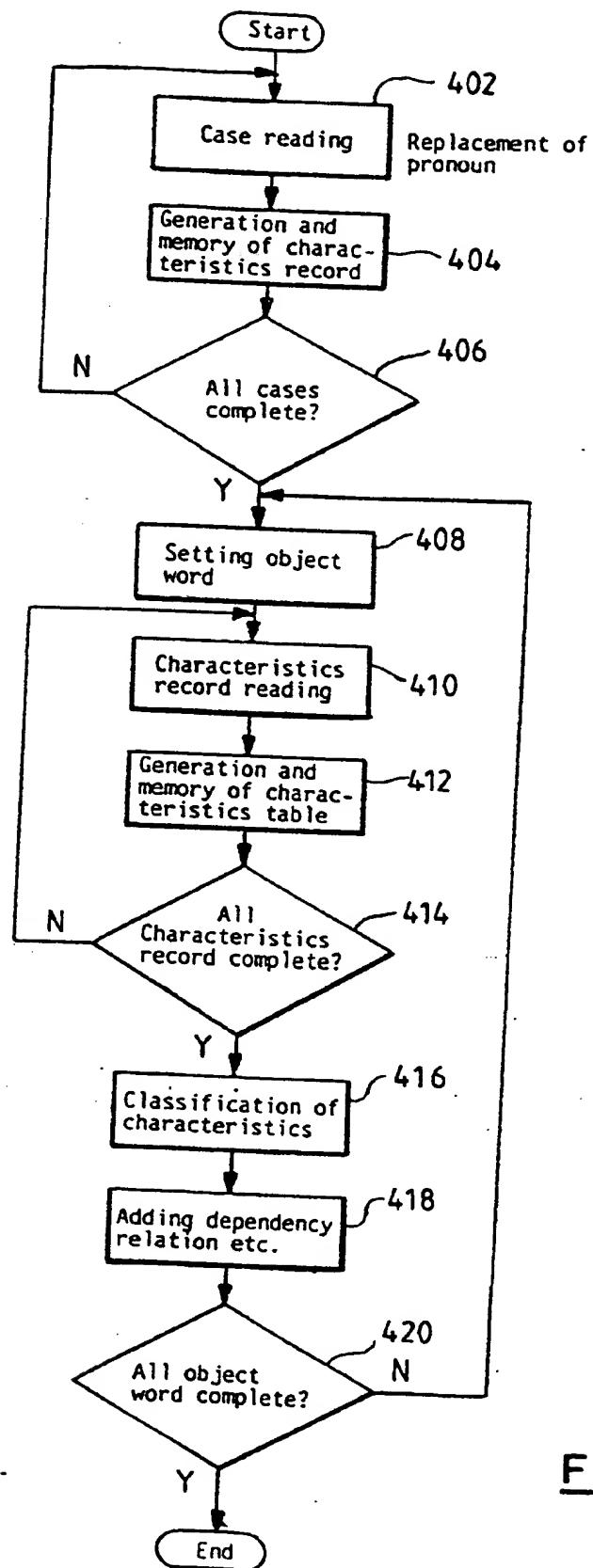
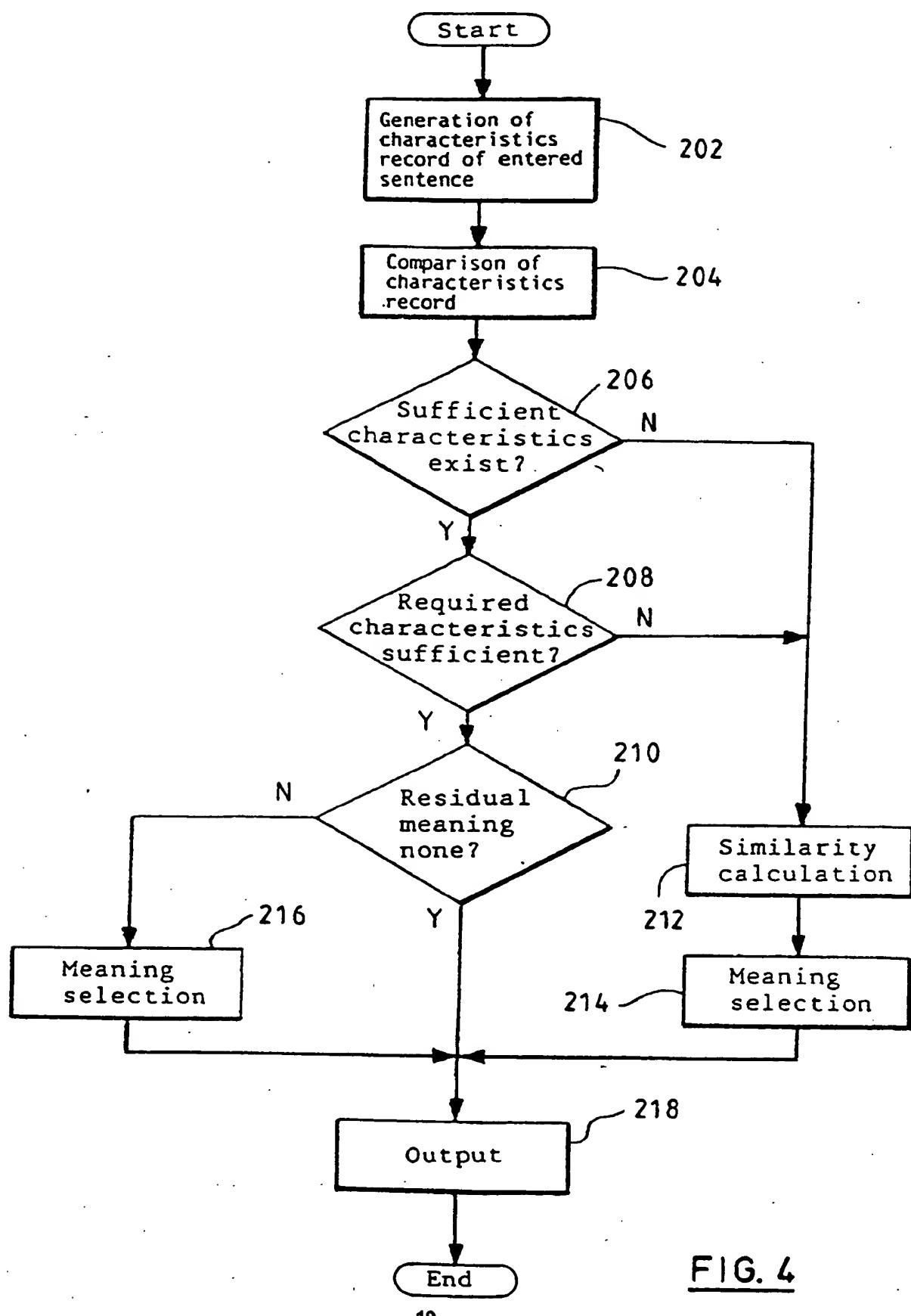


FIG. 2

FIG. 3

FIG. 4



## EUROPEAN SEARCH REPORT

Application Number  
EP 94 30 2567

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cls)
A	US-A-5 181 163 (NAKAJIMA ET AL.) * column 1, line 67 - column 2, line 67 * ---	1-6	G06F15/38
A	IBM JOURNAL OF RESEARCH AND DEVELOPMENT, vol.32, no.2, March 1988, NEW YORK, US pages 185 - 193 E. BLACK 'An experiment in computational discrimination of English word senses' * the whole document *	1-6	
A	US-A-4 942 526 (OKAJIMA ET AL.) * column 3 *	1-6	
A	GB-A-2 220 285 (HITACHI) * page 2, line 1 - page 4, line 5 *	1-6	
A	US-A-4 661 924 (OKAMOTO ET AL.) * column 2, line 12 - line 50 *	1-6	
A	US-A-4 868 750 (KUCERA ET AL.) * column 1, line 60 - column 3, line 49 *	1-6	
A	US-A-4 916 614 (KAJI ET AL.) * column 1, line 63 - column 2, line 58 *	1-6	
A	US-A-4 958 285 (TOMINAGA) * column 1, line 61 - column 2, line 14 *	1-6	
A	EP-A-0 403 660 (KABUSHIKI KAISHA) * page 2, line 12 - page 2, line 12 *	1-6	
A	EP-A-0 266 001 (IBM) * the whole document *	1-6	
A	EP-A-0 467 527 (IBM) * page 4, line 37 - page 5, line 43 *	1-6	
		-/-	
The present search report has been drawn up for all claims			
Place of search	Date of completion of the search	Examiner	
THE HAGUE	9 August 1994	Burö, S.P.	
CATEGORY OF CITED DOCUMENTS			
X : particularly relevant if taken alone	T : theory or principle underlying the invention		
Y : particularly relevant if combined with another document of the same category	E : earlier patent document, but published on, or after the filing date		
A : technological background	D : document cited in the application		
O : non-written disclosure	L : document cited for other reasons		
P : intermediate document	& : member of the same patent family, corresponding document		



European Patent  
Office

## EUROPEAN SEARCH REPORT

Application Number  
EP 94 30 2567

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.)
A	US-A-5 056 021 (AUSBORN) * column 2, line 11 - line 59 * ---	1-6	
A	EP-A-0 409 425 (KIM ET AL.) * column 5, line 26 - column 7, line 54 * ---	1-6	
A	ENTITY-RELATIONSHIP APPROACH: A BRIDGE TO THE USER, 16 November 1988, ROME, IT pages 455 - 472 M.T. PAZIENZA & P. VELARDI 'Using a semantic knowledge base to support a natural language interface to a text database' * page 457 * ---	1-6	
A	US-A-5 132 901 (YOKOGAWA) * column 2, line 27 - column 3, line 18 * -----	1-6	
TECHNICAL FIELDS SEARCHED (Int.Cl.)			
The present search report has been drawn up for all claims			
Place of search	Date of completion of the search	Examiner	
THE HAGUE	9 August 1994	Burö, S.P.	
CATEGORY OF CITED DOCUMENTS			
X : particularly relevant if taken alone	T : theory or principle underlying the invention		
Y : particularly relevant if combined with another document of the same category	E : earlier patent document, but published on, or after the filing date		
A : technological background	D : document cited in the application		
O : non-written disclosure	L : document cited for other reasons		
P : intermediate document	& : member of the same patent family, corresponding document		

EPO FORM 1500/2 (1989)

**THIS PAGE BLANK (USPTO)**